WELL-SWITCHING: example of logistic modeling

Many of the wells used for drinking purposes in Bangladesh and South Asia are contaminated with natural arsenic. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases.

The research team (from US and Bangladesh) measured all the wells and labeled as “safe” if the arsenic level was below 0.5 units in 100mg/l.People using unsafe wells were encouraged to switch to nearby safe wells.

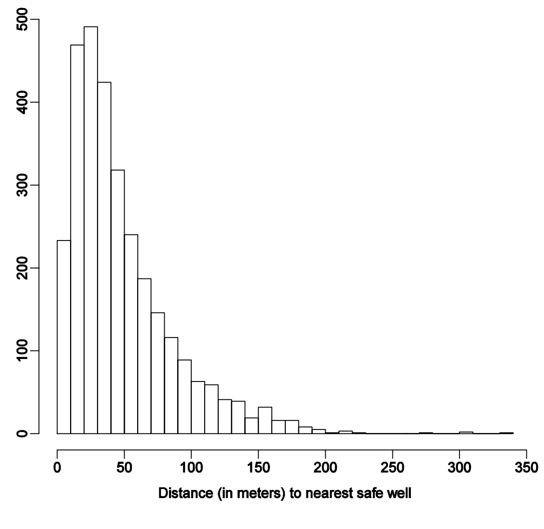
A few years later the research team returned to check who had switched wells to understand the predictive factors for switching. References: Geen, Steen, Vergeng et al (2003) and Gelman, Trevisani,Lu, van Geen, (2004).

THE DATA

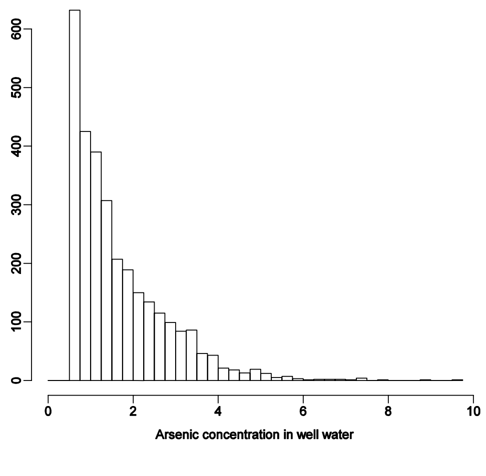
Outcome: y=1, if household switch to a safe well, 0 if not

Predictors:

dist	distance (in meters) to the closest safe well
arsenic	arsenic level of the respondent’s well
educ	education level of the head of the household



Distance



Arsenic level

Both predictors are highly skewed to the left

# Load a program to analyse a logistic regression model

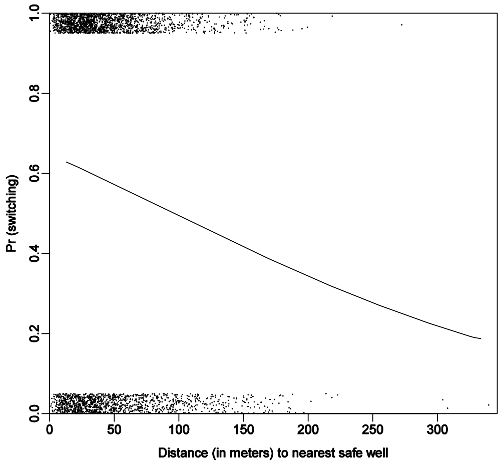
```
>library(arm) # glm module in R (generalized linear models
```

# Model 1. Fit a logistic model for switching given the distance of safe well

```
> fit.1 <- glm (switch ~ dist,family=binomial(link="logit"))
        coef.est coef.se
(Intercept) 0.61 0.06
dist       -0.01 0.00
---
n = 3020, k = 2
residual deviance = 4076.2, null deviance = 4118.1
(difference = 41.9)
```

# Model 2. rescale distance in 100-meter scale

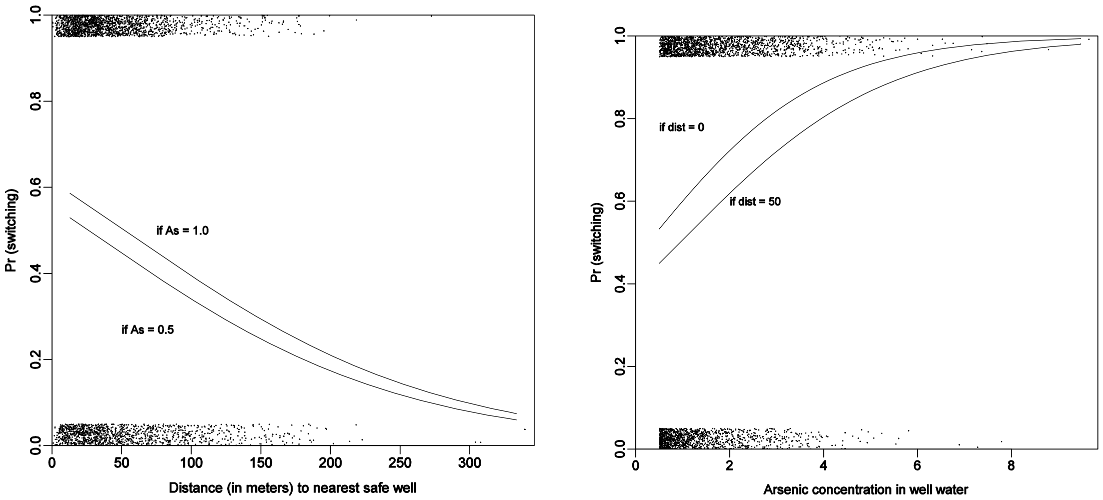
```
> fit.2 <-glm(switch~dist100,family=binomial(link="logit"))
> display (fit.2)
        coef.est coef.se
(Intercept) 0.61 0.06
dist100    -0.62 0.10
---
n = 3020, k = 2
residual deviance = 4076.2, null deviance = 4118.1
(difference = 41.9)
```



Decreasing effect of distance in the logistic model

# Model 3. add arsenic level

```
> fit.3<-
glm(switch~dist100+arsenic,family=binomial(link="logit"))
> display (fit.3)
        coef.est coef.se
(Intercept) 0.00 0.08
dist100    -0.90 0.10
arsenic     0.46 0.04
---
n = 3020, k = 3
residual deviance = 3930.7, null deviance = 4118.1
(difference = 187.4)
```



```
# The longer the distance, the less switching. The higher arsenic level,
the more switching (Note: no linear effect)
```

# Model 4. center the predictors

```
> fit.5 <- glm (switch ~ c.dist100 + c.arsenic +
c.dist100:c.arsenic,family=binomial(link="logit"))
> display(fit.5)
        coef.est coef.se
(Intercept) 0.35 0.04
c.dist100  -0.87 0.10
c.arsenic   0.47 0.04
c.dist100:c.arsenic -0.18 0.10
---
n = 3020, k = 4
```

```
residual deviance = 3927.6, null deviance = 4118.1
(difference = 190.5)
```

# Testing nested models: compare Model 4 vs Model 3 by the difference in deviances vs. the difference in degrees of freedom (k) with the values of the X^2-distribution. Here df is k4-k3=4-3=1, difference in deviance = -2(3927.6-3930.7)= 6.2 and the corresponding X^2=3.841**, p=0.05, so Model 4 is not significantly better than Model 3.**

X^2 distribution: probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

# Model 5. add interaction of distance between arsenic level

```
> fit.7 <- glm (switch ~ c.dist100 + c.arsenic +
c.dist100:c.arsenic +educ4, family=binomial(link="logit"))
> display (fit.7)
        coef.est coef.se
(Intercept) 0.15 0.06
c.dist100  -0.87 0.11
c.arsenic   0.48 0.04
educ4       0.17 0.04
c.dist100:c.arsenic -0.16 0.10
---
n = 3020, k = 5
residual deviance = 3907.9, null deviance = 4118.1
(difference = 210.2)
```

```
# Adding education level of the head of household improves
the fit: difference of deviances 3927.6-3907.9= 26.7 with
only 1 df loss.

# log arsenic level centered
c.log.arsenic <- log.arsenic - mean (log.arsenic)

# Model 6. add interactions of education between distance
and arsenic level

> fit.9 <- glm (switch ~ c.dist100 + c.log.arsenic +
c.educ4 +c.dist100:c.log.arsenic + c.dist100:c.educ4 +
c.log.arsenic:c.educ4,family=binomial(link="logit"))
> display (fit.9)
        coef.est coef.se
(Intercept)   0.35 0.04
c.dist100    -0.98 0.11
c.log.arsenic 0.90 0.07
c.educ4       0.18 0.04
c.dist100:c.log.arsenic -0.16 0.19
c.dist100:c.educ4     0.34 0.11
c.log.arsenic:c.educ4 0.06 0.07
---
n = 3020, k = 7
residual deviance = 3863.1, null deviance = 4118.1
(difference = 255.0)

# No significant interaction effects between arsenic level
and others but the higher education the head of household
has, the more probable to switch even when the distance
increases.


# Average predictive differences: use the model with main
effects only (Model 7.)

fit.10 <- glm (switch ~ dist100 + arsenic + educ4, family =
binomial(link="logit"))
> display (fit.10)
         coef.est coef.se
(Intercept) -0.21 0.09
dist100     -0.90 0.10
arsenic      0.47 0.04
educ4        0.17 0.04
---
n = 3020, k = 4
```

```
residual deviance = 3910.4, null deviance = 4118.1
(difference = 207.7)
```

# compare households with distance next to (lo) and 100 meters from (hi) a safe well

```
> b <- coef (fit.10)
> hi <- 1
> lo <- 0
> delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic +
b[4]*educ4) -
+ invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4)
> print (mean(delta))
[1] -0.2044681
```

# compare households with arsenic level 0.5 (lo) to level 1 (hi)

```
> hi <- 1.0
> lo <- 0.5
> delta <- invlogit (b[1] + b[2]*dist100 + b[3]*hi +
b[4]*educ4) -
+ invlogit (b[1] + b[2]*dist100 + b[3]*lo + b[4]*educ4)
> print (mean(delta))
[1] 0.05643807
```

# compare households with education level 0 (lo) to level 3 (hi)

```
> hi <- 3
> lo <- 0
> delta <- invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*hi)
-
+ invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*lo)
> print (mean(delta))
[1] 0.1167189
```

# Model 8. add interaction between distance and arsenic level

```
> fit.11 <- glm (switch ~ dist100 + arsenic + educ4 +
dist100:arsenic, family=binomial(link="logit"))
> display (fit.11)
          coef.est coef.se
(Intercept)  -0.35 0.13
dist100      -0.60 0.21
arsenic       0.56 0.07
```

```
educ4          0.17 0.04
dist100:arsenic -0.16 0.10
---
n = 3020, k = 5
residual deviance = 3907.9, null deviance = 4118.1
(difference = 210.2)
```

# compare households with distance next to (lo) and 100 meters from (hi) a safe well, now with interaction

```
> b <- coef (fit.11)
> hi <- 1
> lo <- 0
> delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic +
b[4]*educ4 + b[5]*hi*arsenic) -
+ invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4 +
+ b[5]*lo*arsenic)
> print (mean(delta))
[1] -0.1944495
>
```

# Summary: Long distance (100m vs 0m) has a much larger effect on switching than high arsenic level (1 vs 0.5) but if the head of household was educated and aware of the poisonous effect of arsenic, the probability of switching increased even when the nearest safe well was further away.